THE BAYESIAN REVOLUTION IN GENETICS

Mark A. Beaumont and Bruce Rannala‡*

Bayesian statistics allow scientists to easily incorporate prior knowledge into their data analysis. Nonetheless, the sheer amount of computational power that is required for Bayesian statistical analyses has previously limited their use in genetics. These computational constraints have now largely been overcome and the underlying advantages of Bayesian approaches are putting them at the forefront of genetic data analysis in an increasing number of areas.

STATISTICAL INFERENCE
The process whereby data are observed and then statements are made about unknown features of the system that gave rise to the data.

PROBABILISTIC MODEL
A model in which the data are modelled as random variables, the probability distribution of which depends on parameter values. Bayesian models are sometimes called fully probabilistic because the parameter values are also treated as random variables.

LIKELIHOOD
The probability of the data for a particular set of parameter values.

**School of Animal and Microbial Sciences, University of Reading, Whiteknights, P.O. Box 228, Reading RG6 6AJ, UK.*
‡Department of Medical Genetics, 839 Medical Sciences Building, University of Alberta, Edmonton, Alberta T6G2H7, Canada.
Correspondence to M.A.B.
e-mail: m.a.beaumont@reading.ac.uk
doi:10.1038/nrg1318

In many branches of genetics, as in other areas of biology, various complex processes influence the data. Genetics has evolved rich mathematical theories to deal with this complexity. Using these theoretical tools, it is often possible to construct realistic models that explain the data in terms of the processes. Formulating such a model is often the first step towards studying the underlying processes and provides the basis for STATISTICAL INFERENCE. Most genetic properties of individuals, populations or species (such as individual genotypes, population gene frequencies and DNA sequence polymorphisms) are a product of forces that are inherently stochastic and therefore cannot be studied without the use of PROBABILISTIC MODELS. Of course, not every aspect of molecular biology must be studied using probabilistic models. At the biochemical level, for example, particular pathways of gene expression can be studied under more or less controlled conditions that seem (at least to many practitioners) to obviate the need for any statistical analysis. However, even such experimental studies are being increasingly supplemented by the rapidly burgeoning field of functional genomics, a field that has many of the same properties (and problems) as other observational sciences and that requires similar probabilistic analysis.

Genetic data are often the result of a complex process with many mechanisms that can produce the observed data, so what is the best way to to choose among the possible causes? As an example, consider the use of genetic data to identify cryptic population structure (that is, individuals with different population ancestries arising from, for example, geographic separation). The calculation of the chance that an individual carrying a particular genotype was born in a population other than the one from which it is sampled (that is, is an immigrant) depends, among other things, on the gene frequencies in that population. Inferences about the population gene frequencies depend, in turn, on inferences about the populations of origin for all other sampled individuals (given their genotypes), which depend, in turn, on the inferred gene frequencies for all other populations, and so on. Bayesian inference is a convenient way to deal with these sorts of problems (that is, models with many interdependent parameters).

In this review, we compare the Bayesian approach to genetic analysis with approaches that use other statistical frameworks. We endeavour to explain why the use of Bayesian methods has increased in many branches of science during the past decade and highlight the aspects of many genetic problems that make Bayesian reasoning particularly attractive[1]. A potentially attractive feature of Bayesian analysis is the ability to incorporate background information into the specification of the model. However, we argue that the recent popularity of Bayesian methods is largely pragmatic, and can be explained by the relative ease with which complex LIKELIHOOD problems can be tackled by the use of computationally intensive MARKOV CHAIN Monte Carlo (MCMC) techniques. To illustrate this, we describe recent applications of Bayesian inference to three areas of modern genetic analysis: population genetics, genomics and human genetics (primarily gene mapping). Finally, we highlight some of the current problems and limitations of Bayesian inference in genetics and outline potential future applications.

MARKOV CHAIN
A model that is suitable for modelling a sequence of random variables, such as nucleotide base pairs in DNA, in which the probability that a variable assumes any specific value depends only on the value of a specified number of most recent variables that precede it. In an *n*th-order Markov chain, the probability distribution of a variable depends on the *n* preceding observations.

MARGINAL LIKELIHOOD
Also known as the 'prior predictive distribution'. The probability distribution of the data irrespective of the parameter values.

RANDOM VARIABLE
A quantity that might take any of a range of values (discrete or continuous) that cannot be predicted with certainty but only described probabilistically.

JOINT PROBABILITY DISTRIBUTION
The probability distribution of all combinations of two or more random variables.

PRIOR [DISTRIBUTION]
The probability distribution of parameter values before observing the data.

CONDITIONAL DISTRIBUTION
The distribution of one or more random variables when other random variables of a joint probability distribution are fixed at particular values.

POSTERIOR DISTRIBUTION
The conditional distribution of the parameter given the observed data.

POINT ESTIMATE
A summary of the location of a parameter value. In a Bayesian setting, this is generally the mean, mode or median of the posterior distribution.

INTERVAL ESTIMATE
An estimate of the region in which the true parameter value is believed to be located.

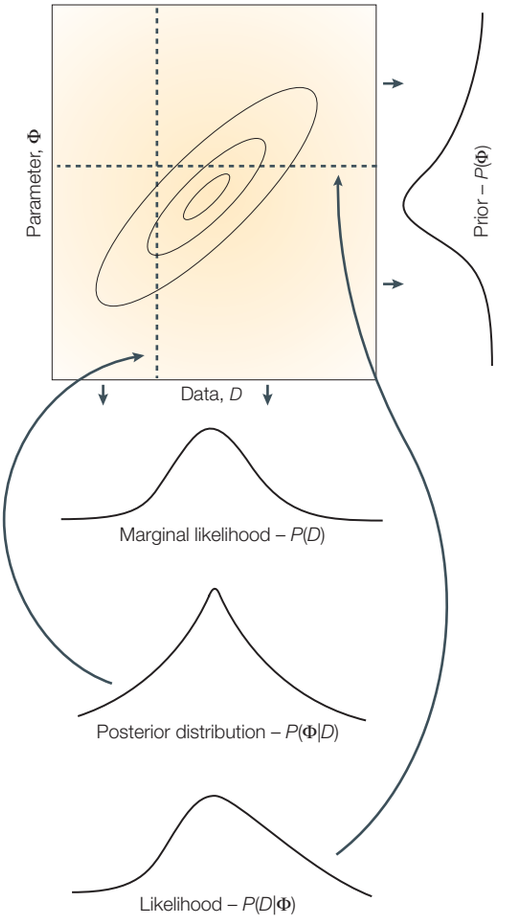


Figure 1 | **The basic features that underlie Bayesian inference.** We imagine that the data *D* can take any value that is measured along the *x*-axis of the figure. Similarly, the parameter value Φ can take any value that is measured along the *y*-axis. Bayesian inference involves creating the joint distribution of parameters and data, $P(D,\Phi)$, illustrated by the contour intervals in the figure. This distribution can be obtained simply as the product of the prior $P(\Phi)$ and the likelihood $P(D|\Phi)$. Typically, the likelihood will arise from a statistical model in which it is necessary to consider how the data can be 'explained' by the parameter(s). The prior is an assumed distribution of the parameter that is obtained from background knowledge. The arrows in the figure show that marginal distributions are obtained by summing (integrating) the joint distribution either over the data, recovering the prior (the distribution on the right of the joint distribution), or over the values of the parameter, giving the MARGINAL LIKELIHOOD (the first distribution directly below the joint distribution). Conditional distributions (represented by the '|' in notation) are indicated by the dotted lines in the figure, and represent taking a 'slice' through the joint distribution and then rescaling the distribution so that the sum (integral) of possible values is equal to one. The scaling factor that is needed is given by the marginal distribution. Any conditional distribution is simply the joint distribution divided by a marginal distribution. For example, the likelihood can be recovered by dividing the joint distribution by the prior. The posterior distribution, $P(\Phi|D)$ — the key quantity that we want in Bayesian inference — is the joint distribution divided by the marginal likelihood. It is the computation of the marginal likelihood (that is, the integrations denoted by the arrows that point down from the joint distribution) that is typically problematic.

Principles of Bayesian inference

The essence of the Bayesian viewpoint is that there is no logical distinction between model parameters and data. Both are RANDOM VARIABLES with a JOINT PROBABILITY DISTRIBUTION that is specified by a probabilistic model. From this viewpoint, 'data' are observed variables and 'parameters' are unobserved variables. The joint distribution is a product of the likelihood and the PRIOR. The prior encapsulates information about the values of a parameter before examining the data in the form of a probability distribution. The likelihood is a CONDITIONAL DISTRIBUTION that specifies the probability of the observed data given any particular values for the parameters and is based on a model of the underlying process. Together, these two functions combine all available information about the parameters. Bayesian statistics simply involves manipulating this joint distribution in various ways to make inferences about the parameters, or the probability model, given the data (FIG. 1). The main aim of Bayesian inference is to calculate the POSTERIOR DISTRIBUTION of the parameters, which is the conditional distribution of parameters given the data.

A POINT ESTIMATE of a parameter is obtained by considering some property of the posterior distribution (usually the mode or the mean). An INTERVAL ESTIMATE of a parameter can be obtained by considering a 'credible set' of values (a set or interval that contains the true parameter with probability $1-\alpha$, for which α is a pre-specified significance level such as 0.05). An example that uses Bayesian inference to 'assign' an individual from an unknown source population to its population of birth on the basis of its genotype is presented in BOX 1.

Other well-known non-Bayesian approaches to statistical inference include the method of maximum likelihood and the METHOD OF MOMENTS, which form the basis of classical or FREQUENTIST INFERENCE[2]. Maximum likelihood bases inferences entirely on the likelihood function, incorporating no prior information and choosing point estimates of parameters that maximize the probability of the data given the parameter (that is, maximizing the likelihood as a function of the parameter for a fixed set of data). Historically, there have been many arguments both for and against the use of various inference frameworks. An old criticism of the Bayesian approach is that there is something unsatisfactorily subjective in choosing a prior. However, this is no different in principle from the choice of likelihood function in the maximum-likelihood method[1]. In fact, as is demonstrated below, modern Bayesian methods often place explicit prior probabilities on alternative likelihood functions to calculate their posterior probability given the data.

There are many practical reasons to use Bayesian inference: if a probability model includes many interdependent variables that are constrained to a particular range of values (as is often the case in genetics), maximum-likelihood inference requires that a constrained multidimensional maximization be carried out to find the combined set of parameter values that maximize the likelihood function. This is often a difficult numerical analysis problem and might require enormous computational effort. In addition, under the maximum-likelihood

Box 1 | **An example of Bayesian inference: assigning individuals to populations**

		Data (observed variables)						
		Genotype *A*			Genotype *B*			
		Likelihood	Joint distribution	Posterior probability	Likelihood	Joint distribution	Posterior probability	Prior probability
Parameters (unobserved variables)	Immigrant	0.01		0.0012	0.99		0.69	
			0.001			0.099		0.1
	Resident	0.95		0.9988	0.05		0.31	
			0.855			0.045		0.9
Probability of data			0.856			0.144		1

This example should be interpreted with reference to FIG. 1**. We imagine a situation in which there are haploid individuals in a population into which immigrants arrive at a low rate. From background information, such as ringing data in birds, we think that the probability that any randomly chosen individual is resident is 0.9 and the probability that it is an immigrant is 0.1: this is our prior (last column on the right). In this population, there are two genotypes at a locus (*A* and *B*). Again from background information, we think that the likelihood of genotype *A* is 0.01 in the immigrant pool and 0.95 in the resident pool (far left column under genotype *A*). The joint distribution is the product of the prior and the likelihood (middle columns under each genotype): this represents the probability of a particular observation. For example, the joint distribution of an immigrant with genotype *A* is 0.001. The probability that an observation will be of a particular genotype, irrespective of whether it is resident or immigrant, is given by the lower margin of the table, which is obtained by summing the joint distribution across parameter values. Given that we observe a particular genotype, the posterior probability that it is either immigrant or resident (right-hand columns under each genotype) is given by the joint distribution scaled so that the sum of possibilities is one, obtained by dividing the joint distribution by the probability of the data. So, if we observe genotype *B*, the posterior probability that it is an immigrant is 0.69 (whereas it was 0.1 before this observation).**

METHOD OF MOMENTS
A method for estimating parameters by using theory to obtain a formula for the expected value of statistics measured from the data as a function of the parameter values to be estimated. The observed values of these statistics are then equated to the expected values. The formula is inverted to obtain an estimate of the parameter.

FREQUENTIST INFERENCE
Statistical inference in which probability is interpreted as the relative frequency of occurrences in an infinite sequence of trials.

COALESCENT THEORY
A theory that describes the genealogy of chromosomes or genes. Under many life-history schemes (discrete generations, overlapping generations, non-random mating, and so on), taking certain limits, the statistical distribution of branch lengths in genealogies follows a simple form. Coalescent theory describes this distribution.

PARAMETRIC BOOTSTRAPPING
The process of repeatedly simulating new data sets with parameters that are inferred from the observed data, and then re-estimating the parameters from these simulated data sets. This process is used to obtain confidence intervals.

EFFECTIVE POPULATION SIZE (N_e)
The size of a random mating population under a simple Fisher–Wright model that has an equivalent rate of inbreeding to that of the observed population, which might have additional complexities such as variable population size or biased sex ratio.

NON-IDENTIFIABLE [PARAMETERS]
One or more model parameters are non-identifiable if different combinations of the parameters generate the same likelihood of the data.

method, calculation of confidence intervals and statistical tests generally involve approximations that are most accurate for large sample sizes — for example, that the probability distribution of the maximum-likelihood estimate follows a normal distribution. On the other hand, in Bayesian inference — in which the prior automatically imposes the parameter constraints — inferences about parameter values on the basis of the posterior distribution usually require integration (for example, calculating means) rather than maximization, and no further approximation is involved. Moreover, numerical methods that were developed in the 1950s using MCMC methods (BOX 2) and implemented on powerful new computers have greatly facilitated the evaluation of Bayesian posterior probabilities, making the calculations tractable for complicated genetic models that have resisted analysis using maximum likelihood or other classical methods. This is arguably the most important factor that drives the recent surge of popularity of Bayesian inference in most branches of science. Here, we present a range of examples in which Bayesian inference has allowed complicated models to be studied and biologically relevant parameters to be estimated, as well as allowing prior information to be efficiently incorporated.

Population genetics

Population genetics has a rich theoretical heritage that stems from the work of Fisher, Haldane and Wright. Initial statistical methods involved calculating expected values of various estimators as functions of parameters in a genetic model and applying the method of moments. Likelihood approaches were not applied to population-genetic problems until later[3,4]. The development of COALESCENT THEORY[5,6] has strongly influenced many areas of population genetics. Similar to earlier approaches, the theory allows the expected values of statistics to be calculated, but also enables sample data sets to be simulated rapidly for PARAMETRIC BOOTSTRAPPING, which in turn allows for more sophisticated calculation of confidence intervals and hypothesis testing in the frequentist tradition. Although not applicable in all areas of population-genetic analysis, the coalescent theory forms the basis for likelihood calculations in genealogical models[7] and has allowed the use of Bayesian approaches to infer demographic history from genetic data (BOX 3). In addition, Bayesian methods have been used to assign individuals to their population of origin and to detect selection acting on genes.

Estimating parameters in demographic models. A feature of population-genetic inference is that parameters in the likelihood function, such as mutation rate (μ) and EFFECTIVE POPULATION SIZE (N_e), occur only as their product (μN_e) — that is, they are NON-IDENTIFIABLE. With non-Bayesian inference, if one parameter is of interest, a 'best-guess' point estimate is typically used for another[8], and there is no rigorous way to incorporate uncertainty. An arguable[9] strength of the Bayesian approach is that prior information can be used to make inferences about non-identifiable parameters[10,11].

Box 2 | Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) describes a class of method that relies on simulating a special type of stochastic process, known as a Markov chain, to study properties of a complicated probability distribution that cannot be easily studied using analytical methods (reviewed in REF. 95). A Markov chain generates a series of random variables such that the probability distribution of future states is completely determined by the current state at any point in the chain. Under certain conditions, a Markov chain will have a 'stationary distribution', meaning that if the chain is iterated for a sufficient period, the states it visits will tend to a specific probability distribution that no longer depends on the iteration number or the initial state of the variable. The basic idea that underlies all MCMC methods is to construct a Markov chain with a stationary distribution that is the probability distribution of interest, and then to sample from this distribution to make inferences. In Bayesian analysis, this distribution is usually the joint posterior distribution of one or more parameters. MCMC has also been used for estimating likelihoods and other purposes in maximum-likelihood inference. Monte Carlo refers to the quarter in the principality of Monaco that is famous for its gambling casinos and alludes to the fact that random numbers are generated to simulate the Markov chain: this method has much in common with generating random events (such as rolling a dice) as is done in games of chance. The simplest form of MCMC is Monte Carlo integration.

Monte Carlo integration

The basic idea that underlies Monte Carlo (MC) integration is that properties of random variables (such as the mean) can be studied by simulating many instances of a variable and analysing the results (reviewed in REF. 96). Each replicate of the MC simulations is independent and the procedure is therefore equivalent to taking repeated samples from a Markov chain that is 'stationary' at points that are sufficiently separated so that they are not correlated. MC integration has been widely applied in statistical genetics (see, for example, REF. 97). The MC simulation method has the advantage that the estimates obtained are unbiased and the standard error of the estimates can be accurately estimated because the simulated random variables are independent and identically distributed. A disadvantage is that with complex multidimensional variables that have a large state space (for example, a range of possible values), enormous numbers of replicate simulations are needed to obtain accurate parameter estimates.

Metropolis–Hastings algorithm

The Metropolis–Hastings (MH) algorithm[98,99] is similar to the MC simulation procedure in that it aims to sample from a stationary Markov chain to simulate observations from a probability distribution. However, in this case, rather than simulating independent observations from the stationary distribution, it simulates sequential values from the chain until it converges and then samples simulated values at intervals from the chain to mimic independent samples from the stationary distribution. The MH algorithm has the advantage that it can improve the efficiency of simulations when the state space is large because it focuses the simulated variables on values with high probability in the stationary chain. Disadvantages include the fact that in most practical applications, there are no rigorous methods available to determine when the chain has converged or what the optimal intervals between samples are to extract the most information while preserving independence between observations.

HIERARCHICAL BAYESIAN MODEL
In a standard Bayesian model, the parameters are drawn from prior distributions, the parameters of which are fixed by the modeller. In a hierarchical model, these parameters, usually referred to as 'hyperparameters', are also free to vary and are themselves drawn from priors, often referred to as 'hyperpriors'. This form of modelling is most useful for data that is composed of exchangeable groups, such as genes, for which the possibility is required that the parameters that describe each group might or might not be the same.

APPROXIMATE BAYESIAN COMPUTATION
The data are simplified by representation as a set of summary statistics and simulations used to draw samples from the joint distribution of parameters and summary statistics (that is, the distribution shown in figure 1). The posterior distribution is approximated by estimating the conditional distribution of parameters in the vicinity of the summary statistics that are measured from the data (the vertical dotted line in figure 1) avoiding the need to calculate a likelihood function.

MULTILOCUS GENOTYPES
The combinations of alleles that are observed when individuals are simultaneously genotyped at two or more genetic marker loci.

ASSOCIATION STUDY
If two or more variables have joint outcomes that are more frequent than would be expected by chance (if the two variables were independent), they are associated. An association study statistically examines patterns of co-occurrence of variables, such as genetic variants and disease phenotypes, to identify factors (genes) that might contribute to disease risk.

INBREEDING COEFFICIENT
The probability of homozygosity by descent — that is, the probability that a zygote obtains copies of the same ancestral gene from both its parents because they are related.

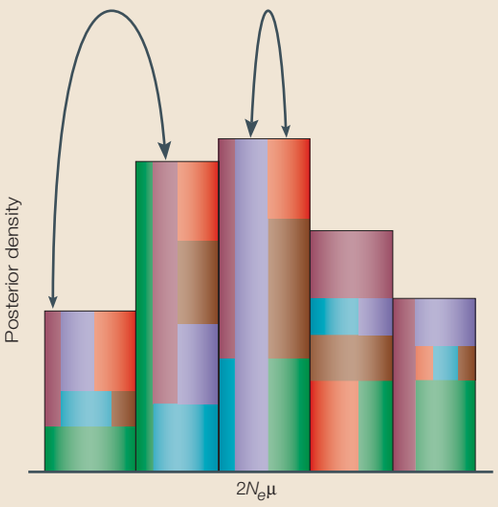
Demographic models often have many parameters and it is conceptually easier to make inferences about them individually, or at most, jointly as pairs. Through the use of marginal posterior distributions, Bayesian analysis deals with this problem simply and flexibly. The classical alternatives are to use point estimates for other parameters or to construct confidence intervals on the basis of profile likelihood[12]. However, in demographic inference, likelihood functions can be complicated and the approximations behind the construction of frequentist confidence intervals are probably not accurate and are technically difficult to apply with a large number of parameters[13,14]. Variability among loci in parameters such as mutation rates can be addressed through the use of HIERARCHICAL BAYESIAN MODELS[15,16] (BOX 4) — for which no classical counterpart is readily available.

As a result of these strengths, Bayesian analysis has in recent years become more prevalent in demographic inference (BOX 5). Computational difficulties can be addressed by improving the efficiency of MCMC methods[16], and also through the use of alternatives to MCMC. An example of the latter is what has come to be known as 'APPROXIMATE BAYESIAN COMPUTATION' (ABC)[17], which in comparisons[18] with the evaluation of the same problem through MCMC[19] can be up to 1,000 times faster, and only slightly less accurate.

Bayesian assignment methods. The study of population differences using genetic markers has a long history (reviewed in Cavalli-Sforza *et al.*[20]). However, it is only relatively recent that methods to assign individuals to populations on the basis of MULTILOCUS GENOTYPES (assignment methods) have been developed. The fundamental equation used in assignment methods calculates the probability of an individual's multilocus genotype given the allele frequencies at different loci in different populations (see BOX 1). The range of practical applications of such assignment tests has proven to be broad. These applications include everything from detecting cryptic population admixture in ASSOCIATION STUDIES[21–24] to detecting population sources of sporadic outbreaks or emerging epidemics[25,26].

Recently, individual assignment methods have been extended in several new directions. Many of these new applications rely heavily on Bayesian methodologies and MCMC techniques. In particular, several new Bayesian methods have been proposed to allow the combined inference of both the partitioning of individuals into subpopulations and the assignment of individual migrant ancestries[27,28]. Another recently proposed method aims to enable the joint inference of the presence of subpopulations within a larger population and the estimation of traditional fixation indices

Box 3 | **Use of MCMC to infer parameters in genealogical models**



Markov chain Monte Carlo (MCMC) methods can be used to obtain posterior distributions for demographic parameters, even though it is only possible to calculate likelihoods for individual genealogies. It is assumed that the parameter of interest is twice the product of the effective population size (N_e) and mutation rate. For simplicity, the prior for any parameter value is a constant, and, therefore, the posterior density for a parameter is proportional to the likelihood. From coalescent theory, we can calculate the probability of the data for a specific parameter value and specific genealogy. The MCMC is assumed to have two types of move: changing the parameter value, keeping to the same genealogy and changing the genealogy, keeping the same parameter value. The moves are reversible but those towards higher likelihoods are favoured (represented by the larger arrow heads in the figure). Relative likelihood is indicated by the area of each individual rectangle. The same genealogy is represented by the same colour. The relative likelihood for particular parameter values is the sum of the relative likelihoods of the genealogies, and provided that a representative sample of genealogies is explored, the MCMC will visit parameter values in proportion to their relative likelihood.

(F statistics[29]) among and within the identified subpopulations[30]. Finally, a Bayesian MCMC method has been proposed for inferring short-term migration rates (over the past few generations) using individual multilocus genotypes[31]. This method also allows for deviations from the Hardy–Weinberg equilibrium (that is, the genotype proportions expected under random mating) within populations by including a separate INBREEDING COEFFICIENT for each population (the value of the inbreeding coefficient is estimated as part of the MCMC inference procedure). The multidimensional complexity of these models makes maximum-likelihood inference difficult and no comparable maximum-likelihood methods have been developed. Multilocus assignment tests are currently in their infancy, but we expect that within a few years they will become a routinely used tool of biologists in fields as disparate as epidemiology, human gene mapping and behavioural ecology.

Detecting selection. Both COMPARATIVE METHODS and population-genetic methods can be used to identify candidate loci that might have been affected by selection[32]. In the case of population-genetic analysis, one idea is to use hierarchical Bayesian demographic models (BOX 4) in which the demographic parameters are allowed to vary among loci to mimic the effects of selection[33,15]. If the posterior probability of zero variance in demographic parameters among loci is itself close to zero, it is probable that some of these loci have been subject to selection. A similar approach has been used to identify candidates for adaptive selection in subdivided populations[34]. A method for finding the distribution of selective effects among loci has also been described[35].

Population-genetic methods for detecting selection might be sensitive to the model that is fitted because demographic events, such as bottlenecks, might mimic or mask the effects of selection[36]. More robust inference is possible using sequence data from different species, in which demographic effects are irrelevant because the segregating variants within a population are not being considered[36]. Analyses at this level focus on the ratio w of nucleotide substitutions that leave the amino acid unchanged in the protein to substitutions that result in a change. If all amino-acid replacing substitutions are neutral, this ratio should be equal to one. If they are deleterious, this ratio should be less than one, and if favoured (positive selection), it should be more than one. Based on these principles, a Bayesian approach has been used to identify which codons are under positive selection in a gene[37]. In this approach (an EMPIRICAL BAYES PROCEDURE), maximum likelihood-generated point estimates of phylogenetic parameters are used to calculate the posterior probability that a codon belongs to one of three categories ($w = 0.1$, or >1). Bayesian phylogenetic methods (see REF. 38) might allow more fully Bayesian estimates of these probabilities.

Genomics

Sequence Analysis. The non-phylogenetic aspects of sequence analysis have a rich and diverse history of model-based methods[39], and include an early application of MCMC to a biological problem[40].

Markov chains or HIDDEN MARKOV MODELS (HMMs) are at the heart of most maximum-likelihood methods of sequence analysis[41]. These methods use DYNAMIC PROGRAMMING to find high-dimensional maximum-likelihood solutions. Some likelihood-based analyses produce scoring functions that involve a Bayesian calculation. For example, the GeneMark software[42], which is used to annotate prokaryote genomes, calculates the likelihood under several different situations (the probability of the data given that it is coding, non-coding, and so on) and then makes an empirical Bayes calculation to pick between them — similar to that described above for detecting nucleotides under selection.

A rich strand of Bayesian analysis has stemmed from models that assume that the bases at nucleotide positions, or amino-acid residues, are drawn at random from frequency distributions that vary among regions. The inference problem is then to locate the regions, marginal to other parameters such as base composition within and outside regions. In this context, Bayesian methods initially were used to model protein alignment[40–43], an approach that has been extended to local alignment[44], and have also been used to identify transcription-factor binding sites[45]. Bayesian modelling based on this approach has been used to obtain the marginal distribution of change points (boundaries of regions) and base compositions along a sequence[46] (see

COMPARATIVE METHODS
Methods for comparing traits across species to identify trends in character evolution that indicate the effects of natural selection.

EMPIRICAL BAYES PROCEDURE
A hierarchical model in which the hyperparameter is not a random variable but is estimated by some other (often classical) means.

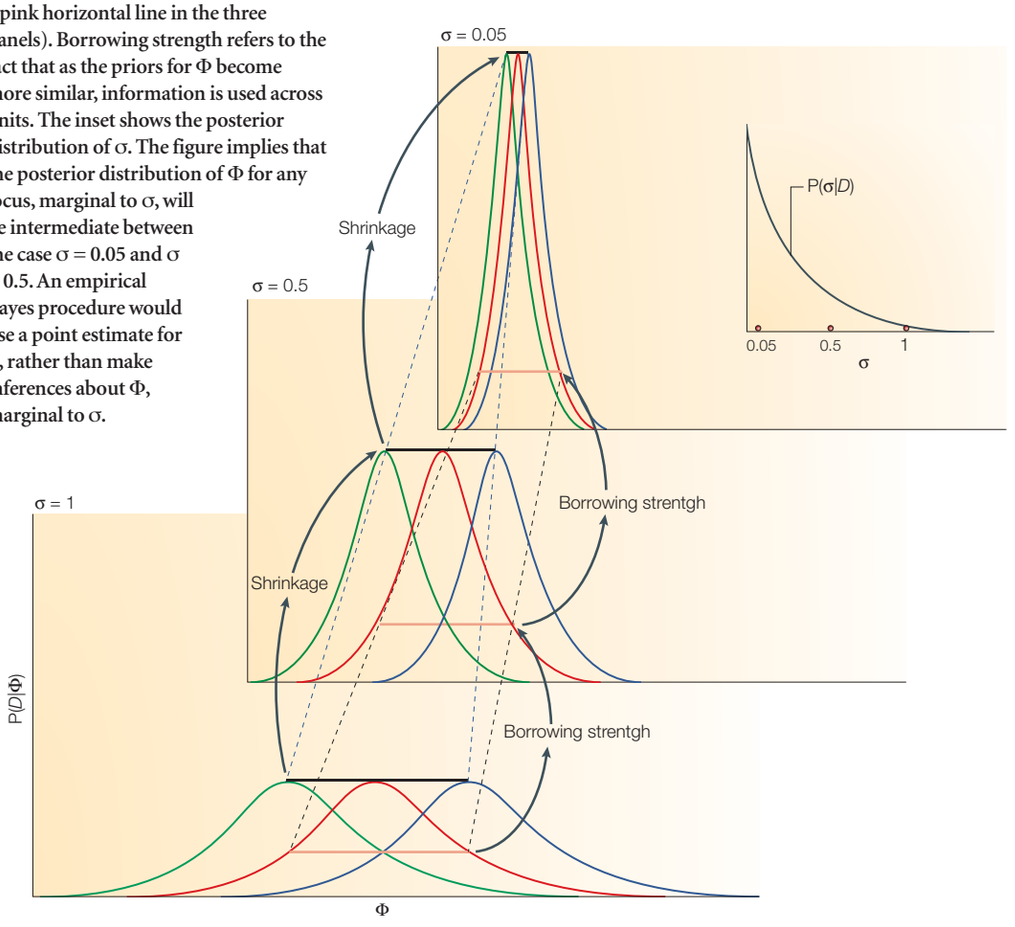
HIDDEN MARKOV MODEL
This is an enhancement of a Markov chain model, in which the state of each observation is drawn randomly from a distribution, the parameters of which follow a Markov chain. For example, the parameter might be an indicator for whether a DNA region is coding or non-coding, and the observation is the base at each nucleotide.

DYNAMIC PROGRAMMING
A large class of programmimg algorithms that are based on breaking a large problem down (if possible) into incremental steps so that, at any given stage, optimal solutions are known sub-problems.

Box 4 | Hierarchical Bayesian models

In a standard Bayesian calculation, as in FIG. 1, the posterior distribution, $P(\Phi|D)$, is proportional to $P(D|\Phi)P(\Phi)$. For example, Φ might be a mutation rate and $P(\Phi)$ might be a prior for the mutation rate. Later, however, it might become apparent that the mutation rate varies among loci, and that there are two causes of uncertainty: uncertainty in the 'type' of locus and uncertainty in the mutation rate given that type. Therefore, rather than combine these two sources of uncertainty into $P(\Phi)$, it is possible to split it into two parts so that σ is a parameter that reflects the type of locus and $P(\Phi|\sigma)$ is the uncertainty in mutation rate given that it is σ. Analagously, Φ might be variance among replicates in expression levels in a microarray experiment. Again, the variance might itself vary among genes, specified by σ. In these cases, Bayesian calculation could be written as $P(D|\Phi)P(\Phi|\sigma)P(\sigma)$. The parameter σ is then often referred to as a 'hyperparameter' and $P(\sigma)$ as a 'hyperprior'.

For data from a single unit, such as a locus, this might not make much difference in the model, depending on how the priors and hyperpriors are specified. However, if the data consist of several different loci, the types of which can be regarded as a random sample from the distribution that is specified by σ, we can then make inferences about σ, as indicated in the figure. The figure shows the posterior distribution of the parameter Φ inferred for three different units (loci/genes), conditional on three different values of the hyperparameter σ that controls variability in Φ among units. As σ becomes smaller (tends to zero; top panel), the posterior distributions of Φ for each unit become more similar, resulting in more similar means (shrinkage; compare the range of means indicated with a black horizontal line in the three panels) and a reduction in variance occurs (BORROWING STRENGTH; compare the variances of the middle distribution indicated with a pink horizontal line in the three panels). Borrowing strength refers to the fact that as the priors for Φ become more similar, information is used across units. The inset shows the posterior distribution of σ. The figure implies that the posterior distribution of Φ for any locus, marginal to σ, will be intermediate between the case $\sigma = 0.05$ and $\sigma = 0.5$. An empirical Bayes procedure would use a point estimate for σ, rather than make inferences about Φ, marginal to σ.



BORROW STRENGTH
This is the tendency in a hierarchical Bayesian model for the posterior distributions of parameters among exchangeable units (for example, genes) to become narrower as a result of pooling information across units.

MODEL SELECTION
The process of choosing among different models given their posterior probability.

also REF. 47). Maximum-likelihood approaches to a problem such as this are generally restricted in the number of parameters considered, and significance testing is often limited because of the high-dimensional optimizations required[46]. By contrast, the Bayesian approach allows more parameters to be considered (essentially allowing parameters that are assumed to be fixed in maximum-likelihood approaches to vary in the Bayesian analysis), it enables full inference on each parameter and allows more rigorous significance testing through MODEL SELECTION. It is often straightforward to incorporate an HMM model into a MCMC framework[48] (see also REF. 47), and so it is likely that Bayesian analyses for sequence data will become more widespread in future, built on the maximum-likelihood framework.

Box 5 | **Examples of Bayesian analysis in demographic inference**

Inferring changes in population size

The first fully Bayesian genealogical analysis was applied to Y-linked microsatellite (YLM) data[11]. Subsequently, there has been interest in inferring population growth. Both approximate Bayesian computation[100] and Markov chain Monte Carlo[19] approaches have been used for YLM data (these approaches yield similar results[18]). Methods for unlinked microsatellite markers have also been developed[33,101].

Analysis of population structure

Models of populations that diverge and evolve independently without gene flow have been considered both for DNA sequence data[16] and also for YLM data[19] — the latter allowing complex bifurcating histories to be considered. A method that enables both migration and population splitting for DNA sequence data has also been developed[13]. Equilibrium models with a constant level of migration between populations seem not to have been directly addressed (but an option for Bayesian analysis is now available in the distributed package for the maximum-likelihood estimation method in REF. 12).

Use of temporal samples

Bayesian methods have been developed to deal with genetic data that are taken at different times, allowing for population growth[102]. This additional temporal information can remove the problem of non-identifiability of parameters. It is then possible to include ancient DNA data to make more accurate inferences about population demography. The method also has applications in viral epidemiology[103]. Furthermore, simpler models can be used to estimate effective population size in the short-term monitoring of populations[104].

Identification of SNPs. The Human Genome Project[49,50] has generated an interest in the identification of nucleotide sites that are polymorphic among individuals — that is single nucleotide polymorphisms (SNPs). There is a large number of SNPs that potentially could be used as markers that are efficient and inexpensive to genotype. The advantages of SNPs for modelling demographic history are offset by the problems of modelling their ascertainment[14,51]. Typically, SNPs are identified by intensively sequencing a small sample of individuals. However, several factors, such as genotyping errors, can lead to a large number of false positives. This presents an ideal problem for Bayesian modelling in which there are data that can be explained by competing hypotheses, but in which we have prior information with which to make judgements among them.

The details of how the Bayesian approach can be applied will obviously depend on the technical details of how the SNPs are identified. A software package that is widely used in non-human[52] as well as human genotyping is PolyBayes[53] (see REF. 54 for a related approach). Two important problems in the identification of SNPs are the presence of PARALOGOUS sequences and sequencing errors. Bayesian calculations can deal with both these issues sequentially[53]. In the first case, the number of mismatches of a sample sequence from a reference sequence is measured. Using prior information on the average pairwise differences between paralogous sequences versus homologous sequences, the probability of obtaining any given number of mismatches under either hypothesis is calculated to obtain the posterior probability that a sequence is not paralogous to the reference sequence. Sequences in which this posterior probability is higher than some critical value are then selected out. The second stage involves performing another Bayesian calculation using aligned sequences, this time with two competing models: first, that the observed variants are the result of sequencing error, and second, that the observed variants are true polymorphisms. In this case, insertions and deletions are ignored. Initial indications are that this is an efficient approach: in a large data set of ESTs, this method discarded around 99.9% of cases as false positives (that is, those in which the variation is inferred to be the result of sequencing error) and 60% of the remaining SNPs were confirmed in a subsequent analysis[53].

Bayesian haplotype inference through population samples. The inference of haplotypes (that is, determining the phase of non-allelic polymorphisms) is an important goal for many reasons (see REFS 55–65). Haplotype phase can be determined in several ways, including linkage analysis[55] and direct molecular techniques, but most are too unreliable, too expensive or too time-consuming to be routinely used. Recently, population-genetic techniques have been proposed for inferring haplotype phase using population samples of genotypes[56–59] based on the principle that the distribution of (observed) multilocus genotypes in a random sample of individuals carries information about the underlying distribution of (unobserved) haplotypes.

Bayesian methods[58,59] have been proposed as an alternative to the Expectation-Maximization (EM) algorithm[60] (a maximum-likelihood approach) for inferring haplotypes from population-genetic data because they do not require all the haplotype frequencies to be retained in computer memory and eliminate the computationally expensive maximization step of the EM algorithm. The Bayesian approach seeks to estimate the posterior probability distributions of the population haplotype frequencies, F, and/or the individual diplotypes (pairs of haplotypes), H, given the sampled genotypes, G. This requires that an explicit prior probability distribution for the population haplotype frequencies, $\Pr(F)$, be specified. Niu *et al.*[58] use an arbitrary distribution for F, whereas Stephens *et al.*[59] use a distribution that is loosely based on a population-genetic (coalescent) model. Although the methods of Stephens *et al.* and Niu *et al.* differ in many of the details, the basic approach is similar.

A shortcoming of current applications of haplotype-inference algorithms is that the resulting haplotypes are often used directly in subsequent studies (for example,

PARALOGOUS
This refers to sequences that have arisen by duplications within a single genome.

ELSTON–STEWART ALGORITHM
An iterative algorithm for linkage mapping. The algorithm calculates the likelihood of marker genotypes on a pedigree. Calculations on the basis of the algorithm are efficient for relatively large families, but its application is typically limited to a small number of markers.

LANDER–GREEN–KRUGYLAK ALGORITHM
An iterative algorithm that is used for linkage mapping. It iteratively calculates the likelihood across markers on a chromosome, rather than across families, as in the Elston–Stewart algorithm. This allows efficient calculation of pedigree likelihoods for small families with many linked markers.

case–control tests for disease–haplotype associations) without accounting for the uncertainty of the individual's inferred haplotypes. In other words, a point estimate of the individual haplotype is treated as an observation in carrying out such tests and this can make the test outcome unreliable if the posterior distribution of haplotypes is not highly concentrated. New methods are needed for carrying out tests of association, and so on, that integrate over the posterior probability distribution of haplotypes and thereby explicitly take account of uncertain phase in carrying out the test. A likelihood ratio test for differences in haplotype frequencies between cases and controls has been proposed by Slatkin and Excoffier[61], but equivalent Bayesian methods have yet to be developed.

Inferring levels of gene expression and regulation. The introduction of methods for measuring levels of gene expression on the basis of DNA/RNA hybridization has provoked substantial interest in the statistical problems that arise[62]. Bayesian statisticians have taken on the challenge of this showcase area in droves, although many of these studies remain in the statistical journals. Although interesting statistical problems are raised in the actual processing of signals from hybridization data[63], the questions that have attracted most attention are: which genes are affected by treatments (for example, tissues and times after treatment, and so on), and what is the model structure that best characterizes expression patterns?

Two issues are important when evaluating the effect of treatment on expression level: making maximum use of the information among genes to model variability among replicate experiments using a particular gene, and minimizing the false-positive and false-negative rates. In the first case, the idea is that with limited replication, it is difficult to be sure whether an observed difference is significant or not; therefore, we need to use the information from other genes. This can be achieved using a hierarchical Bayesian model, in which it is possible to borrow strength from different genes (BOX 4): a partially Bayesian treatment along these lines has already been proposed[64]. These and similar methods would then use a sequential *p*-value method to minimize the number of false positives (for example, see REF. 65). Alternatively, a more fully Bayesian method is possible[66,67], in which the affected genes are picked out through model selection. The advantage of this approach is that great flexibility can be introduced into deciding the level of stringency of discrimination[68].

Microarray studies are often used to group genes that show similar patterns of expression with different treatments. Traditionally, non-parametric ordination or clustering techniques have been used[69]. The advantage of applying Bayesian modelling instead is that it is then possible to carry out statistical tests and obtain confidence bounds on particular groupings, which are not easily obtained using the classical approaches. One approach, which models time-series gene-expression data using regression in a Bayesian framework, defines partitions in which genes have the same regression parameters, and then hierarchically clusters expression patterns on the basis of the posterior probability of partitions, starting with an initial state in which each gene belongs to its own partition[70].

Box 6 | **Analysis of complex traits and quantitative trait locus mapping**

Complex genetic traits, such as body weight or height and many human diseases (for example, type II diabetes and schizophrenia), are determined by the combined influences of multiple genes and the environment. Such polygenic traits are often referred to as 'quantitative' because they are most often measured traits that have a more or less continuous distribution in the population. Genes that have a major effect on a quantitative trait are known as quantitative trait loci (QTLs). A common goal of much research in animal and plant genetics, as well as in human-disease genetics, is to map QTLs to regions of chromosomes in the hope that the causal loci might ultimately be identified by positional cloning. In animal populations, QTL mapping has been carried out for many years using controlled crosses. In humans, controlled crosses are not possible (for obvious reasons) and existing pedigrees must instead be used to map the loci through linkage analysis. Mapping through pedigrees has recently become popular in agricultural and livestock genetics as well.

One serious problem that is encountered when attempting to map QTLs through pedigree analysis is that the QTLs that influence human diseases, or other traits, often have low penetrance (penetrance refers to the probability that an individual who carries one or more copies of the gene has the disease/trait). Low penetrance greatly reduces the power of linkage analysis[55]. The size of the pedigrees can be increased to compensate for this reduction in power. However, maximum-likelihood methods for multipoint linkage analysis that use the ELSTON–STEWART ALGORITHM[105] or the LANDER–GREEN–KRUGYLAK ALGORITHM[106,107] are limited to either a small number of linked loci or fewer than approximately a dozen individuals per pedigree, respectively. Recently, Markov chain Monte Carlo methods for carrying out linkage analysis under complex models of inheritance have been developed[108,109]. The methods seem promising in that they allow much larger pedigrees to be analysed for many linked loci. Several of the most recently developed methods are Bayesian (reviewed by REF. 110) owing to the fact that the complex multidimensional space of the pedigree analysis problem with complex traits has limited progress for maximum-likelihood methods.

Human genetics

The rapid expansion of human genetic data during the past few decades is unprecedented. The Human Genome Project produced a genetic blueprint of our chromosomes[49,50] and documented similarities and differences between individuals; the current haplotype map project (HapMap; see online links box) seeks to further characterize the distribution of nucleotide polymorphisms across chromosomes in human populations[71]. These data present new opportunities to identify genes that are involved in human diseases, for both simple single-gene disorders, such as cystic fibrosis, and complex disorders that are caused by multiple genes and the environment, such as schizophrenia (reviewed in REF. 72; see BOX 6). Genetic marker polymorphisms in human populations can be used to identify genes or genomic regions that are associated with diseases and to aid in the positional cloning of a disease mutation. These objectives require complex statistical modelling, and Bayesian inference has made more rigorous statistical methods feasible in both areas.

Association mapping. Association-mapping methods attempt to locate disease mutations by detecting associations between the incidence of a genetic polymorphism and that of a disease (reviewed in REF. 73). Often referred to as 'case–control studies', such methods have seen widespread application to disease studies using genetic

markers in recent years. Association studies that rely on linkage disequilibrium might provide a new tool for mapping genes that influence complex diseases (reviewed in REF. 74).

Although association methods have been shown to be potentially more powerful than linkage analysis for detecting genes that influence complex disease in some circumstances, they are plagued by false-positive results for various reasons[73]. One source of false-positive associations is population stratification. If a disease mutation and a particular marker allele both happen to have an increased, or decreased, frequency in some particular population (for example, owing to random effects such as joint genetic drift to a higher, or lower, frequency of susceptibility alleles and other non-causal alleles, or as a result of confounding variables such as environmental effects), the allele and the disease might seem to be associated; however, the allele is really a marker of population affiliation rather than being linked to a disease locus and is therefore a false association.

In the early 1990s, FAMILY-BASED ASSOCIATION TESTS (FBATs), such as the transmission disequilibrium test[75], were proposed to allow association studies to be carried out in the presence of population stratification. The basic idea was to examine trios of parents and an affected offspring and to use the non-transmitted alleles from parents as controls and the transmitted alleles as cases. This procedure insures that the proper control allele is used in each comparison even in cases in which the parental mating represents admixture between populations. The currently available FBATs have several shortcomings. First, they test the composite null hypothesis of either no linkage or no association. In many cases, either linkage or association might be of specific interest. Second, the methods do not readily allow information from other prior linkage or association studies to be incorporated into the test. Recently, a Bayesian FBAT has been proposed as a potential solution[76]. The new method combines the likelihood function for FBATs developed by Sham and Curtis[77] with flexible prior probability densities for model parameters such as the recombination fraction between the disease and marker loci that allow either uninformative (uniform) or informative priors to be used depending on the available information. Standard techniques for model testing, based on the BAYES FACTOR, are then used to directly test specific hypotheses about linkage, and so on.

An alternative way to correct for the effects of population stratification in association analyses is to examine unlinked genetic markers (so-called 'genomic controls') to correct for population subdivision in association studies[21]. Multilocus assignment tests developed in recent years[78,79] have been applied to the problem of association mapping in admixed populations[21,22]. These methods have at least two limitations: they were not specifically developed for mapping susceptibility alleles that influence complex traits, and they do not adequately account for the statistical uncertainty of genomic ancestries and admixture proportions. Several Bayesian approaches have been proposed that attempt to correct for these deficiencies. Sillanpaa *et al.*[80] proposed a fully Bayesian approach for association-based quantitative trait locus mapping using unlinked neutral markers as genomic controls. More recently, Hoggart *et al.*[81] proposed a hybrid Bayesian–classical method that uses MCMC to integrate over uncertain admixture proportions and uncertain numbers of founding populations that are involved in an admixture, with a classical generalized linear model approach used to specify trait values.

Fine-mapping of disease-susceptibility genes. In the 1980s, the first genome-wide genetic markers were developed using restriction fragment length polymorphisms (RFLPs). This allowed disease genes to be assigned to specific chromosomal intervals using pedigree-based linkage analysis and raised the possibility of positionally cloning a disease gene. The size of a candidate interval defined by linkage analysis (determined by the number of informative meioses) is typically 1 Mb or more, however, which is much larger than could be sequenced using 1980s technologies. One solution is to genotype polymorphic markers that span the candidate region among unrelated individuals. In this way, 'ancestral' haplotypes that are shared between disease chromosomes can be detected and used to further narrow the candidate region[82,83]. The basic idea is that disease mutations arise on particular chromosomes that carry specific haplotypes, and ancestral recombination increasingly disrupts haplotype sharing in regions that are further from the disease-mutation location[84]. Because alleles at markers near a disease mutation are in greater linkage disequilibrium (LD) than those further away, this technique has come to be known as LD MAPPING.

Early methods for LD mapping could only be used for pairwise analyses using single-linked genetic markers — the basic approach was to solve for the expected fraction of non-recombinant haplotypes under a simple demographic model and then to use this result to derive an estimate of the disease location assuming a Poisson recombination process on the candidate interval[85]. Subsequent methods used parametric models based on coalescent theory that were more realistic for human populations and solved for the maximum-likelihood estimate of the disease-mutation position (reviewed in REF. 86). As the models were made more realistic, and attempts were made to include factors such as multiple linked markers and genetic heterogeneity (for example, multiple disease alleles), it became increasingly difficult to derive tractable maximum-likelihood estimates. Bayesian methods that use MCMC offer a potentially powerful alternative for such analyses. These methods allow integratation (average) over nuisance parameters such as the unknown genealogy (coalescent tree) and ancestral haplotypes that underlie a sample of disease (and control) chromosomes[87,88], and over the unknown ages of disease mutations[89]. These new methods also allow the direct use of multilocus haplotypes or genotypes[90,91] and have been extended to allow the incorporation of additional genomic information into LD mapping through the prior for the disease location. Rannala and Reeve[87] used information from an annotated human genome sequence (National Center for Biotechnology

FAMILY-BASED ASSOCIATION TESTS
A general class of genetic association tests that uses families with one or more affected children as the observations rather than unrelated cases and controls. The analysis treats the allele that is transmitted to (one or more) affected children from each parent as the 'case' and the untransmitted allele is treated as the 'control' to avoid the influence of population subdivision.

BAYES FACTOR
The ratio of the prior probabilities of the null versus the alternative hypotheses over the ratio of the posterior probabilities. This can be interpreted as the relative odds that the hypothesis is true before and after examining the data. If the prior odds are equal, this simplifies to become the likelihood ratio.

LD MAPPING
A procedure for fine-scale localization to a region of a chromosome of a mutation that causes a detectable phenotype (often a disease) by use of linkage disequilibrium between the phenotype that is induced by the mutation and markers that are located near the mutation on the chromosome.

CONVERGENCE
The inexorable tendency for a mathematical function to approach some particular value (or set of values) with increasing *n*. In the case of Markov chain Monte Carlo, *n* is the number of simulation replicates and the values that the chain approaches are the posterior probabilities.

Information (NCBI); see online links box) and the Human Gene Mutation Database (HGMD; see online links box) to modify prior probabilities for the location of a novel disease mutation taking account of the likelihood that disease mutations reside in introns, exons or non-coding DNA. Other innovations made possible by the Bayesian approach include the direct use of genotype data, rather than haplotypes[90,91], by integrating over possible haplotypes in the MCMC algorithm. Allelic heterogeneity can also be modelled using so-called 'shattered coalescent' methods that model independent disease mutations as having separate underlying genealogies[88].

Prospects and caveats

The enormous flexibility of the Bayesian approach, illustrated by the examples given in this article, also points to the need for rigorous model testing. In frequentist inference, a common practice has been to simulate large numbers (thousands) of test data sets in which the true parameter values are known, and then measure the bias, mean squared error and coverage of the estimates. Such a method sits uneasily within the Bayesian model, but is often the simplest way to compare with frequentist approaches[18]. For model-checking in Bayesian inference, it has been suggested that parameters should be drawn from the posterior distribution and then used to simulate other data sets[2]. This is the posterior predictive distribution — the distribution of other data sets given the observed data set. Summary statistics measured in the real data can then be compared with those in the simulated data to see whether the model is reasonable. However, in practice this approach has seldom been taken. Similarly, although it is important to check the sensitivity of the model to the priors, in complicated hierarchical models it is generally unfeasible to systematically examine the effect of different priors on the many parameters in the model. Another issue for studies based on MCMC is the problem of assessing CONVERGENCE, which can be particularly acute for models with a variable number of dimensions. Generally, most Bayesian methods are slow, which provides a strong disincentive for anything more than rudimentary model-checking.

Current trends indicate that modifications to standard MCMC methods will be increasingly explored[92]. For cases in which there are a large number of parameters that are not of interest (such as genealogical history in population-genetic models) and only a few that are of interest, the ABC[18,17] approach seems particularly promising. It is also a 'democratizing' method in that it will attract, for example, biologists, who enjoy computer simulation but have little background in probability, into converting their favourite simulation into a tool for inference. Another burgeoning area, not covered in this review, is the use of Bayesian networks for combining the results from different analyses on the same data sets[93,94]. It could, however, be argued that such approaches, although useful and commercially advantageous, are technical fixes that do not easily lend themselves to scientific enquiry. By contrast, the methods described here are based on probabilistic models of the processes that give rise to a pattern. They have parameters that bear some relation to quantities that could in principle be measured and tested. At the moment, the Bayesian revolution is in its earliest phase, and it will be some time yet before the dust has settled and we can judge which are the most promising avenues for exploration.

1. Shoemaker, J. S., Painter, I. S. & Weir, B. S. Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet.* **15**, 354–358 (1999).
2. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman and Hall, London, 1995).
3. Cavalli-Sforza, L. L. & Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Evolution* **32**, 550–570 (1967).
4. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).
The first use of a sampling distribution in population genetics. This paper anticipates modern approaches, such as the coalescent theory, that model the sampling distribution of chromosomes.
5. Kingman, J. F. C. The coalescent. *Stochastic Process. Appl.* **13**, 235–248 (1982).
6. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
7. Felsenstein, J. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.* **59**, 139–147 (1992).
8. Griffiths, R. C. & Tavaré, S. Ancestral inference in population genetics. *Statistical Sci.* **9**, 307–319 (1994).
9. Markovtsova, L., Marjoram, P. & Tavaré, S. The effect of rate variation on ancestral inference in the coalescent. *Genetics* **156**, 1427–1436 (2000).
10. Tavaré, S., Balding D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
11. Wilson, I. J. & Balding, D. J. Genealogical inference from microsatellite data. *Genetics* **150**, 499–510 (1998).
An early paper that uses MCMC to carry out a fully Bayesian analysis of population-genetic data.
12. Beerli, P. & Felsenstein, J. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl Acad. Sci. USA* **98**, 4563–4568 (2001).
13. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
14. Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.* **69**, 1332–1347 (2001).
15. Storz, J. F., Beaumont, M. A. & Alberts, S. C. Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. *Mol. Biol. Evol.* **19**, 1981–1990 (2002).
16. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656 (2003).
17. Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **100**, 15324–15328 (2003).
18. Beaumont, M. A., Zhang, W., & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
19. Wilson, I. J., Weale, M. E. & Balding, D. J. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. Roy. Stat. Soc. A Sta.* **166**, 155–188 (2003).
20. Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1994).
21. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
22. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
23. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
24. Pritchard, J. K. & Donnelly, P. Case–control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
25. Davies, N., Villablanca, F. X. & Roderick, G. K. Bioinvasions of the medfly *Ceratitis capitata*: source estimation using DNA sequences at multiple intron loci. *Genetics* **153**, 351–360 (1999).
26. Bonizzoni, M. *et al.* Microsatellite analysis of medfly bioinfestations in California. *Mol. Ecol.* **10**, 2515–2524 (2001).
27. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
An influential paper in the development of Bayesian methods to study cryptic population structure. The program described in it, Structure, has been widely used in molecular ecology.
28. Dawson, K. J. & Belkhir, K. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* **78**, 59–77 (2001).
29. Wright, S. *Evolution and the Genetics of Populations: The Theory of Gene Frequencies* (Chicago Univ. Press, Chicago, 1969).
30. Corander, J., Waldmann, P. & Sillanpaa, M. J. Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374 (2003).
31. Wilson, G. A. & Rannala, B. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191 (2003).
32. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nature Rev. Genet.* **4**, 99–111 (2003).
33. Storz, J. F. & Beaumont, M. A. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**, 154–166 (2002).
34. Beaumont, M. A. & Balding, D. J. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* (in the press).
35. Bustamante, C. D., Nielsen, R. & Hartl, D. L. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* **63**, 91–103 (2003).

36. Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647 (2001).
37. Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936 (1998).
The first formal statistical method for inferring site-specific selection on DNA codons.
38. Holder, M. & Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nature Rev. Genet.* **4**, 275–284 (2003).
Reviews the many recent applications of Bayesian inference in phylogeny estimation.
39. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis*, (Cambridge Univ. Press, Cambridge, 1998).
40. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
The methods and models used in this paper have led to the development of a large number of Bayesian methods for the analyses of sequence data by some of the authors and their groups.
41. Churchill, G. A. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94 (1989).
One of the earliest papers to use a hidden Markov model to analyse DNA sequence data.
42. Borodovsky, M., McIninch & J. Genmark: parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123–133 (1993).
43. Liu, J. S., Neuwald, A. F. & Lawrence, C. E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Ass.* **90**, 1156–1170 (1995).
44. Webb, B. M., Liu, J. S. & Lawrence, C. E. BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Res.* **30**, 1268–1277 (2002).
45. Thompson, W., Rouchka, E. C., Lawrence & C. E. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.* **31**, 3580–3585 (2003).
46. Liu, J. S. & Lawrence, C. E. Bayesian inference on biopolymer models. *Bioinformatics* **15**, 38–52 (1999).
47. Liu, J. S. & Logvinenko, T. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. & Cannings, C.) 66–93 (John Wiley and Sons, Chichester, 2003).
48. Churchill, G. A. & Lazareva, B. Bayesian restoration of a hidden Markov chain with aplications to DNA sequencing. *J. Comput. Biol.* **6**, 261–277 (1999).
49. Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
50. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
51. Polanski, A. & Kimmel, M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**, 427–436 (2003).
52. Zhu, Y. L. *et al.* Single-nucleotide polymorphisms in soybean. *Genetics* **163**, 1123–1134 (2003).
53. Marth, G. T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nature Genet.* **23**, 452–456 (1999).
54. Irizarry, K. *et al.* Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genet.* **26**, 233–236 (2000).
55. Ott, J. *Analysis of Human Genetic Linkage* (Johns Hopkins, Baltimore, 1999).
56. Long, J. C., Williams, R. C. & Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810 (1995).
57. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
58. Niu, T., Qin, Z. S., Xu, X. & Liu, J. S. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70**, 157–169 (2002).
59. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
60. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* **B39**, 1–38 (1977).
61. Slatkin, M. & Excoffier, L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity* **76**, 377–383 (1996).
62. Butte, A. The use and analysis of microarray data. *Nature Rev. Genet.* **1**, 951–960 (2002).
63. Huber, W., von Heydebreck, A. & Vingron, M. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. & Cannings, C.) 162–187 (John Wiley and Sons, Chichester, 2003).
64. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519 (2001).
65. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
66. Ibrahim, J. G., Chen, M. H. & Gray, R. J. Bayesian models for gene expression with DNA microarray data. *J. Am. Stat. Ass.* **97**, 88–99 (2002).
67. Ishwaran, H. & Rao, J. S. Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Am. Stat. Ass.* **98**, 438–455 (2003).
68. Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M. & Mallick, B. K. Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19**, 90–97 (2003).
69. Zhang, M. Q. Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.* **9**, 681–688 (2003).
70. Heard, N. A., Holmes, C. C. & Stephens, D. A. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. *Department of Statistics, Imperial College, London* [online], <http://stats.ma.ic.ac.uk/~ccholmes/malaria_clustering.pdf> (2003).
71. Dove, A. Mapping project moves forward despite controversy. *Nature Med.* **12**, 1337 (2002).
72. Rannala, B. Finding genes influencing susceptibility to complex diseases in the post-genome era. *Am. J. Pharmacogenomics* **1**, 203–221 (2001).
73. Sham, P. *Statistics in Human Genetics*, (Oxford Univ. Press, New York, 1998).
74. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
75. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
The first application of a family-based association test. The transmission disequilibrium test has been highly influential and spawned many related approaches.
76. Denham, M. C. & Whittaker, J. C. A Bayesian approach to disease gene location using allelic association. *Biostatistics* **4**, 399–409 (2003).
77. Sham, P. C. & Curtis, D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann. Hum. Genet.* **59**, 323–336 (1995).
78. Paetkau, D., Calvert, W., Stirling, I. & Strobeck, C. Microsatellite analysis of population-structure in Canadian polar bears. *Mol. Ecol.* **4**, 347–354 (1995).
79. Rannala, B. & Mountain, J. L. Detecting immigration by using multilocus genotypes. *Proc. Natl Acad. Sci. USA* **94**, 9197–9201 (1997).
80. Sillanpaa, M. J., Kilpikari, R., Ripatti, S., Onkamo, P. & Uimari, P. Bayesian association mapping for quantitative traits in a mixture of two populations. *Genet. Epidemiol.* **21** (Suppl. 1), S692–S699 (2001).
81. Hoggart, C. J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
82. Bodmer, W. F. Human genetics: the molecular challenge. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 1–13 (1986).
83. Lander, E. S. & Botstein, D. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 49–62 (1986).
84. Dean, M. *et al.* Approaches to localizing disease genes as applied to cystic fibrosis. *Nucleic Acids Res.* **18**, 345–350 (1990).
85. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
86. Rannala, B. & Slatkin, M. Methods for multipoint disease mapping using linkage disequilibrium. *Genet. Epidemiol.* **19** (Suppl. 1), S71–S77 (2000).
A comprehensive review of the various likelihood approximations used in linkage-disequilibrium gene mapping.
87. Rannala, B. & Reeve, J. P. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.* **69**, 159–178 (2001).
The first use of the human genome sequence as an informative prior for Bayesian gene mapping.
88. Morris, A. P., Whittaker, J. C. & Balding, D. J. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* **70**, 686–707 (2002).
89. Rannala, B. & Reeve, J. P. Joint Bayesian estimation of mutation location and age using linkage disequilibrium. *Pac. Symp. Biocomput.* 526–534 (2003).
90. Reeve, J. P. & Rannala, B. DMLE+: Bayesian linkage disequilibrium gene mapping. *Bioinformatics* **18**, 894–895 (2002).
91. Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. & Risch, N. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11**, 1716–1724 (2001).
92. Liu, J. S. *Monte Carlo Methods for Scientific Computing* (Springer, New York, 2001).
93. Pavlovic, V., Garg, A. & Kasif, S. A Bayesian framework for combining gene predictions. *Bioinformatics* **18**, 19–27 (2002).
94. Jansen, R. *et al.* A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453 (2003).
95. Ross, S. M. *Simulation*, (Academic, New York, 1997).
96. Ripley, B. D. *Stochastic Simulation* (Wiley and Sons, New York, 1987).
97. Hudson, R. R. Gene genealogies and the coalescent process. *Oxford Surveys Evol. Biol.* **7**, 1–44 (1990).
98. Metropolis, N. Rosenbluth, A. N., Rosenbluth, M. N., Teller A. H. & Teller, E. Equations of state calculations by fast computing machine. *J. Chem. Phys.* **21**, 1087–1091 (1953).
99. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their application. *Biometrika* **57**, 97–109 (1970).
100. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **116**, 1791–1798 (1999).
The first paper to use an ABC approach to infer population-genetic parameters in a complicated demographic model.
101. Beaumont, M. A. Detecting population expansion and decline using microsatellites. *Genetics* **153**, 2013–2029 (1999).
102. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320 (2002).
103. Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. & Rambaut, A. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol.* **20**, 381–387 (2003).
104. Beaumont, M. A. Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160 (2003).
105. Elston, R. C. & Stewart, J. A general model for the analysis of pedigree data. *Human Heredity* **21**, 523–542 (1971).
106. Lander, E. S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* **84**, 2362–2367 (1987).
107. Krugylak, L., Daly, M. J. & Lander, E. S. Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am. J. Hum. Gen.* **56**, 519–527 (1995).
108. Lange, K. & Sobel, E. A random walk method for computing genetic location scores. *Am. J. Hum. Gen.* **49**, 1320–1334 (1991).
109. Thompson, E. A. in *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (eds Keramidas, E. M. & Kaufman, S. M.) 321–328 (Interface Foundation of North America, Fairfax Station, Virginia, 1991).
110. Hoeschele, I. in *Handbook of Statistical Genetics* (ed. Balding, D. J.) 599–644 (John Wiley and Sons, New York, 2001).
An extensive review of methods used to map quantitative trait loci in humans and other species.

Acknowledgements

We thank the four anonymous referees for their comments. Work on this paper was supported by grants from the Biotechnology and Biological Sciences Research Council and the Natural Environment Research Council to M.A.B., and by grants from the National Institutes of Health and the Canadian Institute of Health Research to B.R.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:
OMIM: http://www.ncbi.nlm.nih.gov/Omim
cystic fibrosis | schizophrenia | type II diabetes

FURTHER INFORMATION

Bayesian haplotyping programs:
http://www.stats.ox.ac.uk/mathgen/software.html;
http://www-personal.umich.edu/~qin
Bayesian population genetics programs and links:
http://evolve.zoo.ox.ac.uk/beast;
http://www.maths.abdn.ac.uk/~ijw;
http://www.rubic.rdg.ac.uk/~mab/software.html
Bayesian sequence analysis web sites:
http://www.wadsworth.org/resnres/bioinfo;
http://www.people.fas.harvard.edu/~junliu/index1.html#Computational_Biology
Detecting selection with comparative data, population genetic analysis: http://abacus.gene.ucl.ac.uk/ziheng/ziheng.html
DMLE+ LD Mapping Program: http://dmle.org
Genetic analysis software links (linkage analysis):
http://linkage.rockefeller.edu/soft
Genetic Software Forum (discussion list): http://rannala.org/gsf
HapMap: http://www.hapmap.org
Human Gene Mutation Database:
http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html
National Center for Biotechnology Information:
http://www.ncbi.nlm.nih.gov
SNP discovery software:
http://www.genome.wustl.edu/groups/informatics/software/polybayes/pages/main.html
Software for sequence annotation:
http://opal.biology.gatech.edu/GeneMark
Structure program (Reference 27):
http://pritch.bsd.uchicago.edu
Access to this interactive links box is free online.